進化型エージェントシステムによる自律学習に関する研究

著者	川上 敬
雑誌名	北海道女子大学短期大学部研究紀要
巻	34
ページ	263-271
発行年	1998
URL	http://id.nii.ac.jp/1136/00000979/

北海道女子大学短期大学部研究紀要　第34号
Bulletin of Hokkaido Women's College No. 34

平成10年3月
March, 1998

進化型エージェントシステムによる自律学習に関する研究

A Study on Autonomous Learning by Evolutionary Agent Systems

川　上　　敬
Takashi　KAWAKAMI

I　は　じ　め　に

一般に，ニューラルネットワークの学習では入力信号に対する正しい出力パターンの対を学習データとする教師付き学習である。しかし自律的エージェントの構築を考慮した場合，ほとんどのケースでは完全な教師信号は得られない。そのため教師信号なしの学習アルゴリズムについても様々な取り組みが行われており，その中の一手法として強化学習が注目されている。強化学習は，システムが決定したアクションの結果として環境から報酬のみを受けとり，適切なアクションパターンを学習するアルゴリズムである。そのため環境が明確にモデリング出来ない問題や動的に環境が変化するような問題に対してもロバストな学習システムを構築する可能性が高く，多くの工学分野に適用可能なアルゴリズムとして期待されている。しかしながら，強化学習は正解が陽に与えられないという点で教師無しの学習であるが，環境からの報酬があらかじめ設定された評価関数により与えられる場合には完全な教師無しとは言えない。この問題点を克服するために Ackley と Littman は ERL（Evolutionary Reinforcement Learning）モデル[(1)]を提案している。このモデルではシステムの動作は一般的な強化学習の枠組みと同様に，内在する行動決定モジュールにより決定され，その結果与えられる評価値に基づき好ましい行動パターンを後天的に学習する。ただし評価値は環境から陽には与えられず，環境情報を入力として同様に内在する評価モジュールにより決定される。この評価モジュールは先天的に与えられるため，正しい評価モジュールを持つシステムのみが行動決定モジュールを学習により正しい方向に組織化できる。ERL モデルでは，行動決定モジュールと評価モジュールは供にニューラルネットワークで構成され，その初期状態は遺伝的コードを翻訳する事により生成される。このアプローチは，生物においては評価モジュールが遺伝的に与えられ，行動決定モジュールは内部評価に従い後天的学習を行うという見解に基づいたものといえる。

筆者らはこれまで，強化学習の実現手法として Holland により提案されたクラシファイアシステム[(2)]を利用した学習システムの構築を行い，その有効性を検証してきた[(3)]が，タスクに対する適切な報酬関数設定の問題が常に存在していた。この設定が学習性能に大きな影響を与えるにもかかわらず，経験則や試行錯誤により設定されているのが現状である。

そこで本報告では報酬値を自律的に創出する強化信号生成モジュールを内部に持つクラシファイアシステムモデルを対象とし，この強化信号生成モジュールを進化的に合成するための

手法について述べる。そのために，ERL モデルと同様に遺伝的アルゴリズム(GA)[4] を適用し，良好な強化信号生成モジュールの進化的獲得を提案する。

II 関 連 研 究

ERL モデルは進化機構によってシステムの行動のみならず，そのシステムに内在する目標をも洗練させる機構を有する。すなわち，内在する目標がシステムの学習方向を決定することになる。このメカニズムを実現するために，行動決定ネットワークと評価ネットワークを要素としてシステム内部に持ち，各々はニューラルネットワークで構成され，その初期状態は遺伝的コードを翻訳する事により生成される。センサ入力に従い行動決定ネットワークから何らかの行動が出力され，実行に移される。その結果システムの状態が更新されるが，この状態の好ましさを評価ネットワークが算出する。そして提示された評価値が高くなるように行動決定ネットワークの重み係数が修正される。したがって評価ネットワークはシステムに遺伝的に与えられた目的仮説として機能する。また ERL モデルでは行動決定ネットワークを学習するために CRBP (Complementary Reinforcement Back Propagation) アルゴリズム(1)を提案している。CRBP アルゴリズムでは，評価ネットワークから出力される時刻 t の評価値 E^t と時刻 $t+1$ の評価値 E^{t+1} との差に従って誤差逆伝搬を行うアルゴリズムである。また，北野は ERL モデル包含するような遺伝的監視理論[5]を提案している。ここでは，入力パターンの重要度に基づいた反復学習や遺伝的コードにニューラルネットワークの構造自体を表現する手法等のアイデアが盛り込まれている。これらの研究は真に自律的な適応システムを構築するための有力な方法論であると考えられる。

III アプローチ

ERL モデルではニューラルネットワークを中心とした刺激－反応系を構成しているが，本報告ではクラシファイアシステムに基づく強化学習メカニズムを対象とする。一般にクラシファイアシステムは相互作用し合うサブシステムを構成要素として持ち，プロダクションルールベースと類似の特徴を持つ。ある環境に対する戦略はクラシファイアと呼ばれる if/then 形式のストリングルールによって表現され，このクラシファイアの有限個の集合であるルールベースによりシステムの振る舞いが制御される。実行システムでは刺激－反応系の中心部分となるプロセスが実行される。すなわち，クラシファイアシステムの基本実行サイクルから見ると，以下の 4 つのプロセスが実行される。

1）環境情報のエンコードプロセス：クラシファイアシステムの知覚機関としてのディテクターを通して，環境情報がシステムが処理可能な形式の内部メッセージにコード化される。

2）条件マッチングプロセス：1）で得られたコード化された環境情報とクラシファイア集合中の条件部分との照合により環境にマッチする部分集合を抽出する。

3）ルール競合解決プロセス：条件にマッチしたクラシファイア間の競合により，活性化するクラシファイアが選択される。この競合は各クラシファイアに対応する強度を基にして行われ，より高い強度を有するルールが高い確率で選ばれる。

4）アクションコードのデコードプロセス：3）で選択されたクラシファイアの行為部分をエージェントが実行可能なように記号列から実行命令形式に翻訳する。

各クラシファイアには，環境への適応度に応じて増減される強度と呼ばれる値が割当てられ，システムは学習を重ねるごとにこの強度値を更新し，クラシファイア群を問題に適応させる。この強度値の更新は強化モジュールにより生成された強化信号により行う。しかしインプリメントに際して，報酬関数がタスクに対して適切に設定されなければ，正しい強化信号が導出できないという問題はやはり存在する。この問題に対しても筆者らは報酬関数の各評価項目に対する重み付け係数をGAによりチューニングするアプローチを提案し，その効果を確認した[6]が，評価項目自体が適切に設定できない問題環境の場合には解決策とはならない。したがって，より自律的な適応エージェントを構築するために，エージェントが有するセンサ値のみを入力値として報酬値が得られる機構が必要となる。そこで本報告では，クラシファイアの強度値を更新するための強化信号を生成するモジュールをフィードフォワード型のニューラルネットワークで構築するモデルを提案する（図1）。

強化信号生成モジュールへの入力は各センサ値で，出力としてその状態の好ましさが表現される。従ってシステムはこの評価値の変化に基づいて強度値を更新する。一般的なクラシファイアシステムにおける強度の変更処理は，あるアクションを実行した結果，環境から与えられる報酬rに基づき，そのアクションに関与したクラシファイアの強度を次式に従い更新する。

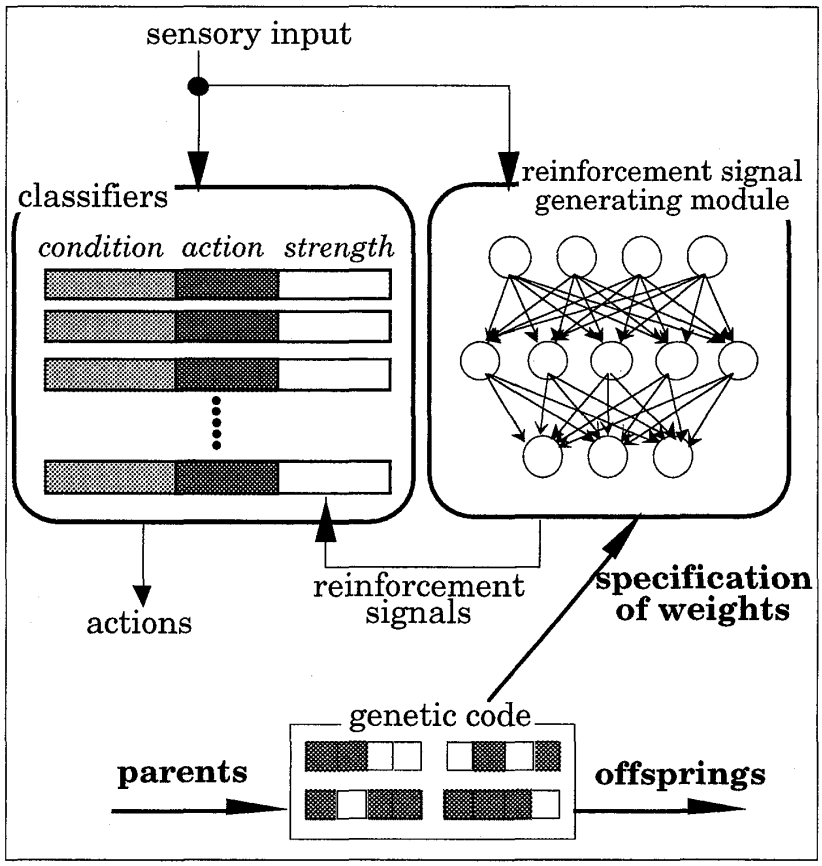

fig.1. Overview of our approach

$$Scf_A{}^{t+1}=Scf_A{}^{t}+Rf(r^t) \tag{1}$$

ここで $Scf_A{}^t$ はアクションに関与したクラシファイア cf_A の時刻 t における強度値で，RF は報酬を強化信号に変換する関数，r^t はその時に与えられる報酬である。

しかし本報告では環境からの報酬値 r^t が明確に設定できない問題環境を扱うためシステム内に存在する強化信号生成モジュールからの評価値を用いて強度更新を行う。従って上式(1)は次のように書き換えられる。

$$Scf_A{}^{t+1}=Scf_A{}^{t}+(E^{t+1}-E^t) \tag{2}$$

ここで E^t はアクションを実行する前の状態に関する評価値，E^{t+1} はアクションを実行した後の変化した状態に関する評価値で，供に強化信号生成モジュールのニューラルネットワークからの出力として与えられる。

この強化信号生成モジュールの構造はERLと同様に遺伝的に決定されリンクの重み等は固定である。従って，問題に対して不適切な構造の強化信号生成モジュールを有するシステムは正しい学習を行うことが出来ない。そこで本アプローチでは，好ましい強化信号生成モジュールの進化的獲得をGAにより試みる。強化信号生成モジュールの進化サイクルは，ニューラルネットワークの構造最適化問題にGAを適用する単純な手法と同様に，1組のウエイトセットを1つの染色体に対応させ，その適応度に基づき遺伝的オペレーションを作用させるものとした。具体的には以下の手順で実行される。

手順1　1組のウエイトセットを1つの染色体として表現し，その初期集団を生成。

手順2　各染色体を翻訳し，個々の強化信号生成モジュールを持つクラシファイアシステムの初期状態を生成。

手順3　各初期クラシファイアシステムをそれぞれ同じタスクにより学習する。

手順4　学習を行った各クラシファイアシステムを学習性能に基づき適応度を算出する。

手順5　適応度に従い，遺伝的オペレータを適用し，新しい染色体集団を生成する。

手順6　手順2）以降を終了条件を満たすまで繰り返す。

Ⅳ　計算機実験

4・1　対象問題

本提案手法を検証するために，ロボットマニピュレータのモーションプランニングタスクを効率的に学習するようなクラシファイアシステム内の強化信号生成モジュールの進化的獲得実験を行う。ロボットマニピュレータのパスプランニング問題とはロボティクスの分野で多くの研究が行われてきた代表的な問題である。与えられた領域内でマニピュレータの初期姿勢から目標姿勢までの動作系列を計画するもので，問題の複雑さによって問題解決の難易度が大きく異なる。たとえば作業空間内の障害物が静的でかつ位置や形状情報が既知である場合には，計

算コストを除けば有用なアルゴリズムが提案されているが，障害物が動的であったり作業空間が未知の場合には，解の導出は非常に困難となる。ここではモデル化したロボットマニピュレータによるタスクを与え，クラシファイアシステムにより学習する。すなわち，マニピュレータの物理的状態を単純化しマニピュレータのリンクは線分として，関節は点としたモデルを採用する。この時，未知作業空間内でのプランニングを対象とし，プランニングエージェントには障害物などの作業環境情報は事前には与えられない。また動作制御は単位時間ステップごとの各関節角度の変位により与え，幾何的情報のみによりプランニングを行う。したがってトルクや慣性等の動力学的要素は無視し，幾何的情報のみによりプランニングを行う。また各関節は回転関節とし，各リンクのねじれ角は全て0とすることによりマニピュレータの作業空間を2次元平面内に限定する。このマニピュレータの姿勢 q は各関節角度によるコンフィギュレーション空間で表現され，関節変数ベクトルで示される。すなわち，

$$q=(q_1,q_2\cdots,q_i\cdots\cdots,q_n)^T. \tag{3}$$

ここで q_i は関節変数で T は転置記号である。また各関節変数 q_i の可動範囲は［0，2π］に設定する。

マニピュレータの動作制御は離散時間ごとの各関節角度変位ベクトル Δq で与える。

$$\Delta q=(\Delta q_1,\Delta q_2,\cdots,\Delta q_i,\cdots,\Delta q_n)^T. \tag{4}$$

したがって時刻 t の姿勢 $q(t)$ は次式で表現される。

$$q(t)=q(t-1)+\Delta q(t). \tag{5}$$

ただしここでは単位時間内における角変位の最大値をあらかじめ与えるものとする。

$$-\alpha\leq\Delta q_i\leq\alpha. \tag{6}$$

次にエージェントが利用できるセンサ情報は，各関節角度変数 q_i と各リンクと障害物との接触情報のみとし，障害物までの距離情報や接触位置情報は知覚できない。したがって，環境内に存在する障害物の位置・形状情報などは与えられない。

上記問題設定にもとづいて，問題の初期状態から1単位時間ごとに1つの動作（これを1モーションと呼ぶ）を繰り返し，目標地点にエンドエフェクターが到達するまでのプランニングが行われ，そのときの角変位ベクトル $\Delta q(t)$ の系列を解とする。1回のプランニングは初期状態から初めて，次のいずれかの場合に終了する。このサイクルを1トライアルと呼ぶ。

1）エンドエフェクターが目標地点に到達した場合。

2）マニピュレータが障害物に接触した場合。

3）予め設定した最大モーション回数に達した場合。

4・2　クラシファイアシステムのインプリメント

本対象問題にクラシファイアシステムをインプリメントする場合，システムが入力とする環境情報にはマニピュレータの関節変数ベクトル q が相当する。また衝突検出センサについては，障害物と接触した時点で，1回のトライアルが終了する。従って関節変数ベクトル q を変換し，各クラシファイアの条件部分とマッチングされる。そして選ばれたクラシファイアの行為部分をデコードすることにより各関節の変位量ベクトル Δq が指示される。その結果，状態が更新され新しい環境情報が観測される。このときクラシファイアとの条件マッチングが行えるように観測情報をストリングへとコード化する必要がある。すなわち連続値で与えられる関節変数 q_i のとりうる範囲を幾つかの領域に分割し，それぞれを1つのコードに割り当てる。ここでは変数領域を8分割し，それぞれを3ビットのコードで表現する。ただし通常の2進数表現ではなく，すべての隣り合った領域同士のハミング距離が1となるようにグレイコーディングを採用する。またクラシファイアの行為部で示される関節角度変位量 Δq_i も同様に8分割しグレイコード化する。

ルールの強化則については，先に提案したようにニューラルネットワークを中心として構築される強化信号生成モジュールから出力される信号値によりクラシファイアの強度を更新する。すなわち本問題の場合，時刻 t における状態のセンサ値として関節角ベクトル $q(t)$ が状態評価ネットワークに入力されその状態の好ましさ E^t が出力される。この時に選択されたクラシファイア $cf_A{}^t$ が示すモーション $\Delta q(t)$ を実行する事によりマニピュレータの状態が $q(t+1)$ に変化する。この新しい状態に対しても同様に評価値 E^{t+1} が与えられるため，この状態遷移に関与したクラシファイア $cf_A{}^t$ の強度値に $E^{t+1}-E^t$ が加算される（図2）。

4・3　状態評価ネットワークの進化的獲得

各クラシファイアに対して強化信号を出力する状態評価ネットワークの構造はクラシファイアシステムが学習を行っている間固定であるため，問題に対して適切な状態評価ネットワークの構造を有さないシステムは正しい学習を行う事が出来ない。そこでここではGAにより，好ましい状態評価ネットワークのウエイトセットを進化的に獲得する。この進化プロセスでは単純GAを用い，使用するGAパラメータとして，染色体の集団数を20，再生はエリート戦略とルーレット型選択戦略を組合わせたもの，交叉確率は0.5，突然

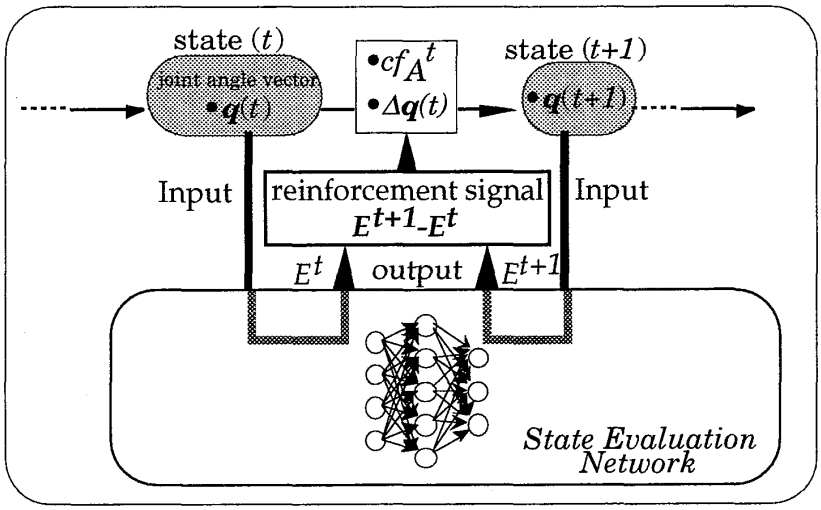

fig.2.　Reinforcement strategy of classifier rules

変異確率は0.3にそれぞれ設定した。

4・4　実験結果

GAによる強化信号生成モジュールの進化的獲得実験を行った。図3は与えた4リンクマニピュレータのモーションプランニングタスクとGAを200世代適用した後，獲得された良好な強化信号生成モジュールを有するクラシファイアシステムにより学習されたプランを示してい

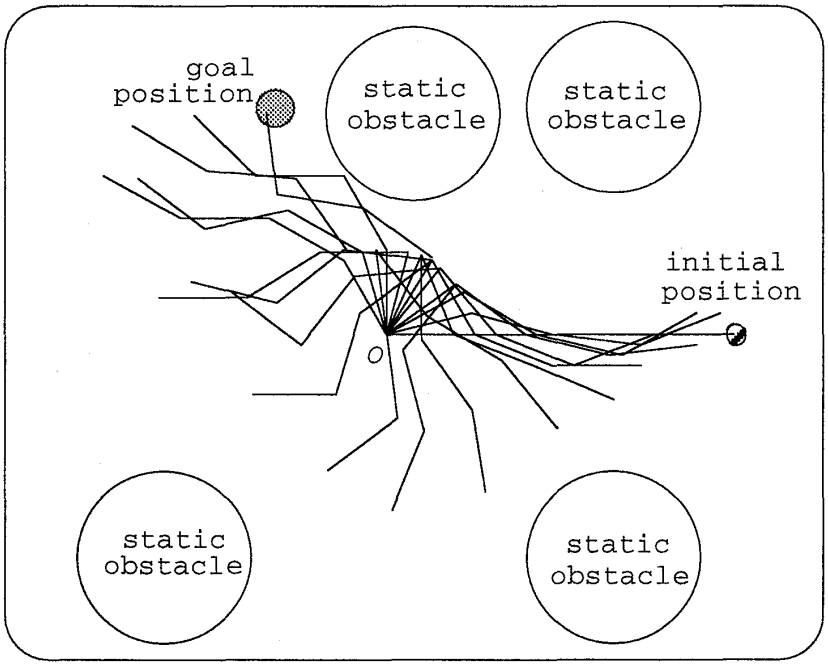

fig.3. The planned motion at the 300th trial. This plan is learned by the classifier system having an approapriate reinforcement signal generating module acquired through 200 generation of genetic algorithms

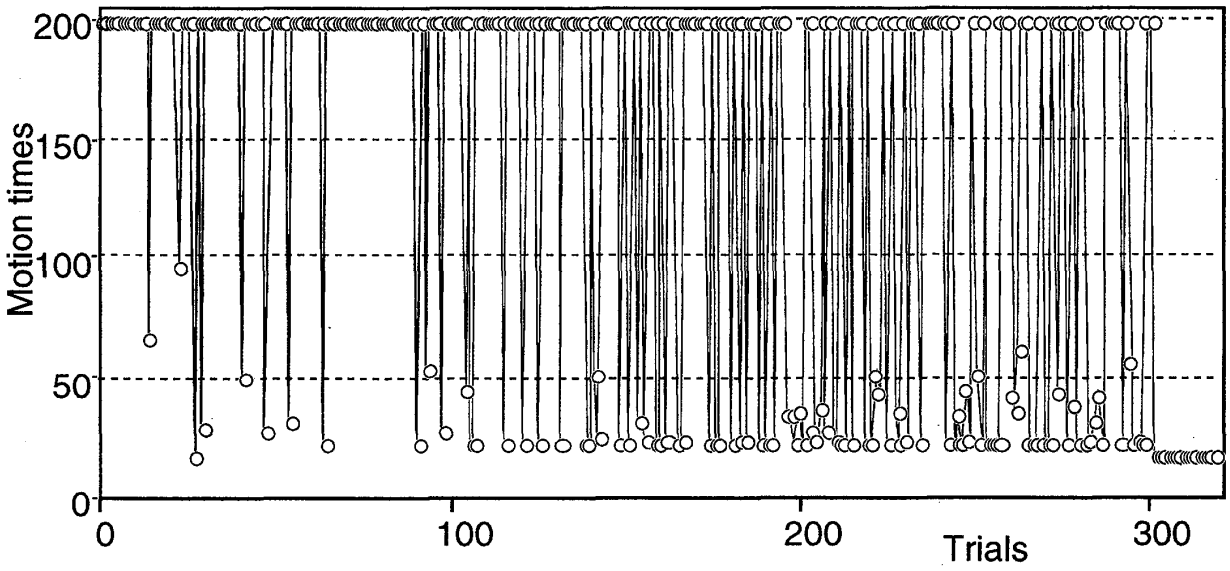

fig.4. A learning curve of 4-link manipulator experiment

る。またこの時の学習曲線を図4に示す。図中縦軸は各学習トライアルにおけるモーション数を示し，モーション数が200回を超えてもゴールに達しない場合は，その時点でそのトライアルを終了させるものとする。

V おわりに

本稿では，広範なタスククラスにおいて，自律型適応エージェントをコントロールできるように，クラシファイアシステムの設計手法を一般化するために，GAによる強化信号生成モジュールの進化的獲得手法を提案した。また通常は評価関数により定義される強化信号をニューラルネットワークにより生成するクラシファイアシステムのモデルを提案した。これにより，より複雑な状態パターンの評価を柔軟に表現できるに効果があると考えられる。さらにロボットマニピュレータのモーションプランニング問題等に本アプローチを適用し，その特性を確認した。

付記

本研究は北海道女子大学短期大学部における平成9年度特別研究費の助成をうけて実施されたものである。

引用・参考文献

(1) Ackley, D. and Littman, M., Interactions Betwen Learning and Evolution, Artificial Life II, Addison-Wesley, (1992), 487.

(2) Holland, J. H., et al., Induction, MIT Press, (1986), 102.

(3) 川上・嘉数，クラシファイアーシステムによる自律型ロボットナビゲーション問題に関する研究，機論，59-564，C (1993)，2339.

(4) Holland, J. H., Adaptation in Natural and Artificial Systems, Univ. of Michigan Press, (1975), 20.

(5) 北野，自律適応システムにおける状況評価回路の進化的獲得，遺伝的アルゴリズム2，産業図書 (1995)，176.

(6) 川上・嘉数，クラシファイアーシステムアーキテクチャの進化的合成に関する研究，機論，62-598，C (1996)，2484.

(7) Grefenstette, J. J., Credit Assignment in Rule Discovery Systems based on Genetic Algorithms, Machine Learning, 3, (1988), 225.

(8) Holland, J. H., Properties of the Bucket Brigade Algorithm, 1st ICGA, (1985), 1.

(9) Wilson, S. W. and Goldberg, D. E., A Critical Review of Classifier Systems, 3rd ICGA, (1989), 244.

(10) Riolo, R. L., Lookahead Planning and Latent Learning in a Classifier Systems, From

Animals to Animats, MIT, (1991), 316.

(11) Unemi, T., et al., Evolutionary Differentiation of Learning Abilities, Artificial Life IV, MIT press, (1994), 331.

(12) Schuurmans, D. and Schaeffer, J., Representational Difficulties with Classifier Systems, 3rd ICGA, (1989), 328.